Generative models of part-structured 3D objects

Charlie Nash
School of Informatics,
University of Edinburgh, UK
charlie.nash@ed.ac.uk

Christopher K. I. Williams
School of Informatics
University of Edinburgh, UK
Alan Turing Institute, London, UK
ckiw@inf.ed.ac.uk

Abstract

We introduce two generative models of part-segmented 3D objects: the shape variational auto-encoder (ShapeVAE) and the shape factor analyzer (ShapeFA). These models describe a distribution over the co-existence of object parts, as well as over the continuous variability of the object surface, leveraging the part structure of 3D objects in their architecture. We demonstrate that while the ShapeFA slightly outperforms the ShapeVAE in terms of density estimation, the ShapeVAE produces better quality samples and is effective at completing partially obscured shapes.

1 Introduction

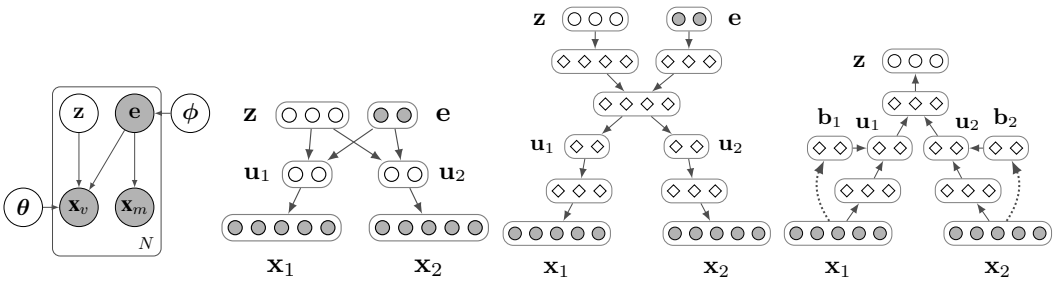
There has been a recent focus on the use of fine-grained shape representations in scene understanding tasks see e.g. [13, 10, 2]. A detailed representation of object shape allows for complex 3D reasoning, and a model of shape variability aids the performance of recognition tasks in images. Structures such as 3D bounding boxes [9, 8], wireframe models [13], or 3D CAD models [2] have been used as shape representations and successfully recognized in images. However, even these more sophisticated object representations are limited in the extent to which they can capture 3D shape. We make use of the representation described by Huang *et al*. [4] consisting of a collection of dense keypoints, segmented into the object's constituent parts (Figure 2). Such a representation is useful as a means of representing 3D structures, and as such a model of the variability of an object class in terms of this representation would be useful for computer vision tasks.

In this work we present two models of structural and local shape variability that capture a distribution over the co-existence of object parts, as well as over the continuous variability of the object surface. We demonstrate that a part-structured variational auto-encoder achieves comparable performance in density estimation to a linear baseline, while producing samples of a higher quality.

2 Model

We consider a dataset of aligned 3D objects from the same object class, with a shape representation consisting of 3D point clouds in which points are in one-to-one correspondence across different instances. We assume that the 3D point clouds consist of multiple parts, which may or not be present in a particular instance as in [4]. In order to deal with missing parts we allow our keypoint variables to take on a symbol m indicating missingness. As such the keypoint variables have the domain $\mathbf{x} \in (\mathbb{R} \cup m)^D$ where D is the number of keypoints. We represent the pattern of part existences in a data example with a binary vector $\mathbf{e} \in \{0, 1\}^K$ where K is the number of object parts, and aim to model the joint distribution over keypoints and part existences $p(\mathbf{x}, \mathbf{e})$.

Part existences: Part existence variables $\mathbf{e}$ are assumed to have been generated by a categorical distribution $p(\mathbf{e}|\boldsymbol{\phi})$. For an object with K parts this is a distribution over 2^K possible states, which in the absence of independence assumptions requires $2^K - 1$ parameters. In practice only a few



(a) Graphical model (b) ShapeFA (c) ShapeVAE decoder (d) ShapeVAE encoder

Figure 1: **Generative models of part-segmented shapes:** (a) Graphical model representing conditional independence assumptions. (b) ShapeFA generative model structure. (c) ShapeVAE generative model structure. Filled circles represent visible variable, un-filled circles represent latent variables, and diamonds represent deterministic variables.

arrangements of part existences occur in object datasets, and so a maximum likelihood estimate assigns most part combinations zero probability.

Keypoints: Keypoints $\mathbf{x}$ are generated by conditioning on part existences $\mathbf{e}$. Let $k(i)$ be the part index associated with keypoint i, and let $\mathcal{M}(\mathbf{e}) = \{i : e_{k(i)} = 0\}$ be the set of indices of missing variables. We can then write the set of missing keypoints as $\mathbf{x}_m = \{x_i\}_{i \in \mathcal{M}}$ and the set of visible keypoints as $\mathbf{x}_v = \{x_i\}_{i \notin \mathcal{M}}$. The missing keypoints $\mathbf{x}_m$ are deterministically mapped to the missing data symbol m, and keypoints whose parts are present $\mathbf{x}_v$ are generated by a Gaussian latent variable model

$$p(\mathbf{x}|\mathbf{e}, \boldsymbol{\theta}) = p(\mathbf{x}_m|\mathbf{e})p(\mathbf{x}_v|\mathbf{e}, \boldsymbol{\theta}) \tag{1}$$

$$= \mathbb{I}[\mathbf{x}_m = [m, \ldots, m]] \int_{\mathbf{z}} \mathcal{N}(\mathbf{x}_v|\boldsymbol{\mu}(\mathbf{z}, \mathbf{e}, \boldsymbol{\theta}), \boldsymbol{\Sigma}(\mathbf{z}, \mathbf{e}, \boldsymbol{\theta}))\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})d\mathbf{z}, \tag{2}$$

where the mappings from latent variables and existences to parameters $\boldsymbol{\mu}(\mathbf{z}, \mathbf{e}, \boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\mathbf{z}, \mathbf{e}, \boldsymbol{\theta})$ can take a variety of forms as shown below. The graphical model representing the joint distribution of keypoints and existence variables is shown in Figure 1a.

Shape factor analysis: In this variant of the keypoint distribution we use a hierarchical factor analysis model. For each part k we introduce part latent variables $\mathbf{u}_k$ and use a factor analysis model to capture shape variability within a particular part. Writing $\mathbf{x}_k$ for the keypoints associated with part k we have:

$$p(\mathbf{x}_k|\mathbf{u}_k, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_k|\mathbf{W}_k^{(1)}\mathbf{u}_k + \boldsymbol{\mu}_k^{(1)}, \boldsymbol{\Psi}_k^{(1)}) = \text{FA}(\mathbf{x}_k|\mathbf{u}_k, \boldsymbol{\theta}_k^{(1)}) \tag{3}$$

For a given set of part existences a conditional distribution over all the visible keypoints given the part latent variables can be obtained by concatenating the part latent variables $\mathbf{u}_v = \{\mathbf{u}_i\}_{i:e_i=1}$ and parameters of the visible parts $\boldsymbol{\theta}_v = \{\boldsymbol{\theta}_i\}_{i:e_i=1}$. A top-level factor analysis model captures the joint distribution of the part latent variables for each part arrangement: $p(\mathbf{u}_v|\mathbf{z}, \mathbf{e}, \boldsymbol{\theta}) = \text{FA}(\mathbf{u}_v|\mathbf{z}, \mathbf{e}, \boldsymbol{\theta})$. By integrating out the part latent variables we obtain a shallow factor analysis model, thus determining the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ functions described in the previous section: $\boldsymbol{\mu}(\mathbf{z}, \mathbf{e}, \boldsymbol{\theta}) = \mathbf{W}_v^{(1)}(\mathbf{W}_\mathbf{e}^{(2)}\mathbf{z} + \boldsymbol{\mu}_\mathbf{e}^{(2)}) + \boldsymbol{\mu}_v^{(1)}$ and $\boldsymbol{\Sigma}(\mathbf{z}, \mathbf{e}, \boldsymbol{\theta}) = \boldsymbol{\Psi}_v^{(1)} + \mathbf{W}_v^{(1)}\boldsymbol{\Psi}_\mathbf{e}^{(2)}\mathbf{W}_v^{(1)\mathsf{T}}$. This structure is illustrated in Figure 1b. We train the model using the greedy layer-wise procedure described by Tang *et al*. [11].

Shape variational auto-encoder: The shape factor analysis model makes strong assumptions about the data distribution: it assumes the data is uni-modal for each part arrangement, and can be generated by adding noise to a linear mapping from latent variables. The decoder of a variational auto-encoder (VAE) [7] is a more flexible class of generative model in which the mapping from latent variables to data variables is an arbitrary neural network (MLP). As such the VAE can capture non-linear structure and multi-modality which may be present in the data.

We take advantage of the part-structure of the data and mimic the architecture of the ShapeFA. As shown in Figure 1c the generative network takes latent variables and existence variables as input, and passes them through a series of fully-connected layers before combining to form a part representation

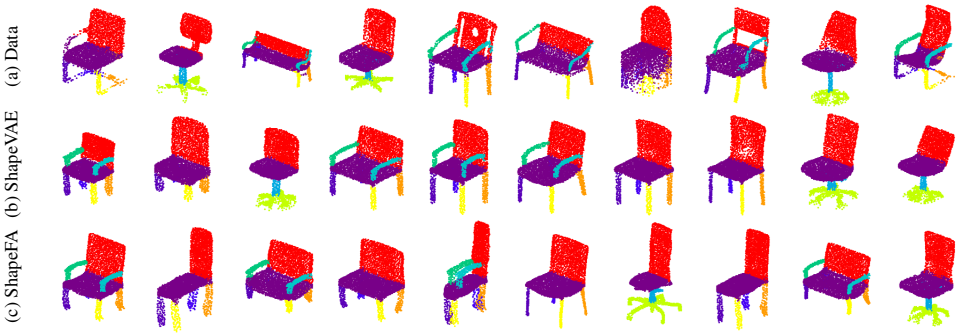


Figure 2: **Shape samples**. (a) A collection of samples from the chairs dataset. (b) Samples generated by a ShapeVAE with 64 latent dimensions. (c) Samples generated by a ShapeFA with 32 latent dimensions.

$\mathbf{u}(\mathbf{z}, \mathbf{e}) = \texttt{MLP}(\mathbf{z}, \mathbf{e})$. This representation is then split into its constituent parts $\mathbf{u} = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$ and passed through further non-linear layers $\mathbf{h}_k(\mathbf{z}, \mathbf{e}) = \texttt{MLP}(\mathbf{u}_k(\mathbf{z}, \mathbf{e}))$. Finally the output parameters are obtained using a linear layer for the mean $\boldsymbol{\mu}_k(\mathbf{z}, \mathbf{e}) = \texttt{Linear}(\mathbf{h}_k(\mathbf{z}, \mathbf{e}))$, and by applying an exponential non-linearity for the variance $\boldsymbol{\sigma}_k^2(\mathbf{z}, \mathbf{e}) = \exp(\texttt{Linear}(\mathbf{h}_k(\mathbf{z}, \mathbf{e})))$. This is analogous to the ShapeFA in which structural latent variables map to latent part variables, which are then mapped independently to their associated keypoints.

The encoder of the ShapeVAE reverses the architecture of the decoder but is modified to make use of information about which parts are missing. As shown in Figure 1d the encoder takes keypoints $\mathbf{x}$ as input and maps to a part representation $\mathbf{u} = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$. For parts that are present the input is mapped through fully connected layers to the part representation, whereas parts that are missing simply generate a bias which is added in the appropriate position to the the part representation:

$$\mathbf{u}_k(\mathbf{x}) = \begin{cases} \texttt{MLP}(\mathbf{x}_k), & \text{if } \mathbf{x}_k \notin \mathcal{M} \\ \mathbf{b}_k, & \text{if } \mathbf{x}_k \in \mathcal{M} \end{cases} \tag{4}$$

The part representation is then passed through fully-connected layers to obtain the output parameters of the approximate Gaussian posterior distribution as in the decoder. The parameters of the encoder and decoder are learned simultaneously using the auto-encoding variational Bayes algorithm [7].

3 Related work

Generative models of 3D objects have been proposed for a range of shape representations including 3D voxel images [12], keypoints [4] and meshes [5, 14]. The closest work to ours is Huang *et al*. [4] in which part-segmented 3D keypoints are modelled with the Beta Shape Machine (BSM), a variant of a multi-layer Boltzmann machine that captures global and local shape variation with a similar part-oriented structure. This model is demonstrated to be effective at generating plausible shapes, as well as for shape segmentation and fine grained classification tasks. Unlike the BSM, our models are directed, and as such training and sampling is more straightforward. In related work Zuffi et al. [14] develop a parts-based 'stitched puppet' model of human shape which allows for the shape and pose of body parts to be modelled separately, while encouraging connecting parts to be close together. This allows for shape variation to be captured on various levels: on the global level articulated pose is modelled, and on the local level continuous shape deformation is modelled

There has been recent work on generative models of voxel representations of 3D objects. Wu et al. [12] model the joint distribution of voxels and object class labels with a convolutional deep belief network. The authors use the model to recognise object classes and reconstruct 3D objects from a single depth image. Girdhar *et al*. [3] use a 3D convolutional auto-encoder to establish a compressed vector representation of 3D objects that can be predicted and reconstructed in real images. Volumetric representations have the advantage that different objects are directly comparable on the voxel level, whereas vertices and faces in triangulated meshes are not comparable. However high resolution voxel grids have very many dimensions which is problematic for efficient generative modelling. For this reason most work so far has made use of quite coarse voxel grids, which limit the fidelity of the objects they can represent.

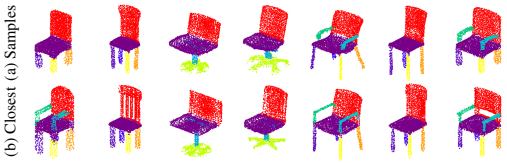


Figure 3: **Generalisation**. *Top:* Shape samples from a ShapeVAE with 64 latent dimensions. *Bottom:* Closest examples in the training set.

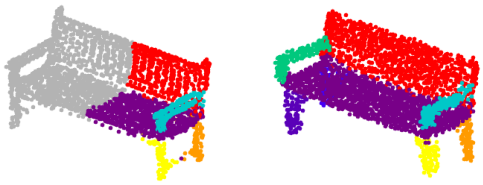


Figure 4: **Shape completion**. *Left:* Unseen data example with half of its variables obscured. *Right:* Shape completion using the ShapeVAE.

Diag-Gaussian	ShapeFA-8	ShapeFA-32	ShapeVAE-8	ShapeVAE-32
0.37 ± 0.32	$\mathbf{1.40 \pm 1.00}$	1.35 ± 1.35	1.35 ± 0.51	1.35 ± 0.35

Table 1: **Test log-likelihood.** Test log-likelihood with standard deviation in nats per dimension for an independent Gaussian baseline, the ShapeFA, and the ShapeVAE.

4 Experiments

Data: We use the chairs class of the dataset developed by Huang *et al.* [4]. This dataset consists of part-segmented 3D point-clouds of aligned objects. The chairs dataset has 3382 examples and demonstrates considerable variability both within parts and in terms of the global layout of parts, and as such it is a challenging task for generative modelling. We thank the authors of [4] for providing us with the high quality part-segmented keypoint data.

We trained ShapeVAE and ShapeFA models with part representations $\mathbf{u}_k$ with $1/15$ of the dimensions of each part $\mathbf{x}_k$. *Hyperbolic tangent* non-linearities were used and the ShapeVAE was trained with the ADAM optimizer [6] with a batch size of 256, and a learning rate of 0.001 for 1200 epochs.

Density estimation: We evaluate the log-likelihood obtained by the ShapeFA and the estimated log-likelihood for the ShapeVAE on a test set for models with 8 and 32 latent dimensions. The ShapeVAE log-likelihood is estimated using 10,000 importance weighted samples [1]. We also include a simple baseline model consisting of a diagonal Gaussian distribution for each existence combination. Table 1 shows that the ShapeFA with 8 latent dimensions achieves the best performance, however comparing with the ShapeVAE-32, the difference in score is not significant at the 5% level under a paired t-test.

Sample quality: We demonstrate features of the generative models by comparing sampled 3D objects to data examples in Figure 2. The ShapeFA produces samples that are mainly plausible, however there are examples of unusually stretched parts. The samples produces by the ShapeVAE are realistic and do not suffer from the same stretching issues as the ShapeFA samples.

Generalisation: As an assessment of ShapeVAE's generalisation capability we show model samples along with their closest training examples in terms of Euclidean distance in Figure 3. The figure shows that for a random selection of samples the nearest data example is qualitatively different in terms of local shape variability. This indicates that the model is generalising beyond the examples seen in the training set.

Shape completion: The ShapeVAE and ShapeFA can be used for the completion of obscured shapes. Figure 4 shows an example of a ShapeVAE completion, which indicates that the model is capable of plausible conditional inference. In this example, one side of the object is shown to the model, and the hidden side is filled in by optimizing the latent variables with respect to the visible variables, and then mapping back to data space using the optimized latent variables.

5 Conclusion

In this paper we have demonstrated that part-structured generative models can effectively model the high dimensional distribution over object shapes, producing realistic and novel samples. In the future we plan to investigate the use of such models as priors for shape estimation tasks with real images.

Acknowledgements

The work of CW is supported in part by EPSRC grant EP/N510129/1. CN is supported by the Centre for Doctoral Training in Data Science, funded by EPSRC grant EP/L016427/1.

References

[1] Y. Burda, R. B. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *CoRR*, abs/1509.00519, 2015.

[2] C. B. Choy, M. Stark, S. Corbett-Davies, and S. Savarese. Enriching object detection with 2D-3D registration and continuous viewpoint estimation. In *CVPR*, pages 2512–2520. IEEE Computer Society, 2015.

[3] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016.

[4] H. Huang, E. Kalogerakis, and B. M. Marlin. Analysis and synthesis of 3D shape families via deep-learned generative models of surfaces. *Computer Graphics Forum*, 34(5):25–38, 2015.

[5] E. Kalogerakis, S. Chaudhuri, D. Koller, and V. Koltun. A probabilistic model for component-based shape synthesis. *ACM Trans. Graph.*, 31(4):55:1–55:11, 2012.

[6] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations*, 2015.

[7] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations*, 2014.

[8] J. Liebelt and C. Schmid. Multi-view object class detection with a 3D geometric model. In *CVPR*, pages 1688–1695. IEEE Computer Society, 2010.

[9] N. Payet and S. Todorovic. From contours to 3D object detection and pose estimation. In *ICCV*, pages 983–990. IEEE Computer Society, 2011.

[10] H. Su, Q. Huang, N. J. Mitra, Y. Li, and L. J. Guibas. Estimating image depth using shape collections. *ACM Trans. Graph.*, 33(4):37:1–37:11, 2014.

[11] Y. Tang, R. Salakhutdinov, and G. E. Hinton. Deep mixtures of factor analysers. In *ICML*. icml.cc / Omnipress, 2012.

[12] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920. IEEE Computer Society, 2015.

[13] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3D representations for object recognition and modelling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2608–2623, 2013.

[14] S. Zuffi and M. J. Black. The stitched puppet: A graphical model of 3D human shape and pose. In *CVPR*, pages 3537–3546. IEEE Computer Society, 2015.